OK or not OK? Commitments in acknowledgments and corrections

For many researchers ([C,G,T], *inter alia*), an acknowledgment as in (1)c by 0 of a discourse move m by 1 can signal that 0 has understood what 1 has said, or that 0 has committed that 1 has committed to a content p with m, and serve to "ground" or to establish a mutual belief that 1 has committed to p. Corrections, and self-corrections, as in (1)d, on the other hand, serve to remove commitments.

(1) a. 0: Did you have a bank account in this bank? b. 1: No sir.
c. 0: OK. So you're saying that you did not have a bank account at Credit Suisse?
d. 1: No. sorry, in fact, I had an account there. e. 0: OK thank you.

The problem is that grounding doesn't follow just from the simple gloss above. Common commitments are needed and don't follow from a simple semantics for acknowledgments. Further, no work has looked at a logical analysis of corrections. We provide such an analysis, showing that these moves have an essential, strategic role to play in dialogue, even if we assume a perfect communication channel and unambiguous commitments in dialogue moves.

We formulate the semantics of dialogue moves and conversational goals in terms of nested, public commitments for reasons given in [V] (*contra* [LA]). Public commitment is an operator with a weak modal logic (K); a player commits to a proposition φ ($C_i\varphi$) given a discourse move m, when m entails φ or when i says φ. In general commitments do not validate type 4 axioms of modal logics; saying φ is not the same as saying *I commit to* φ. We define common commitments for a group G, $C^*_G\varphi$, as $C_G\varphi \wedge C_GC_G\varphi \wedge \dots C_G(C_G)^n\varphi \wedge \dots$ (analogously to common knowledge). Common commitments could follow naturally from assuming: (a) a perfect communication channel and (b) a view of semantic competence that entails perfect knowledge of speaker commitments of unambiguous discourse moves. But then, as in our second dialogue semantics (below), grounding acknowledgments are semantically superfluous: if m entails p, then i's making m entails $C^*_GC_ip$. i's acknowledgment of j can thus only mean that i agrees with the content of j's move, which manifestly it does not, as in (1)c (such acknowledgments are often present in legal questioning). Our first version makes grounding impossible in finite conversations: if a discourse move m by i entails only C_ip, (a) and (b) entail that all the conversational participants believe C_ip [T, G]. Then j's acknowledgment of m would entail $C_jC_ip \wedge Bel_GC_jC_ip$; but using the game theoretic framework of [AP], we show that common commitments are achieved only after an infinite sequence of acknowledgment moves between i and j. Our proposal is that a particular sort of acknowledgment and confirming question licenses the move to common commitment. It is the one in (1)c, where 0 asks a confirming question after an acknowledgment of a move m. If 1's answer to the confirming question is consonant with m, then $C^*_{\{0,1\}}C_1\varphi$, and 0 has achieved her goal.

Can we do without common commitments? We think not; common commitments are essential (see also [C]) for strategic reasons and can be present even when mutual beliefs about a shared task are not. Suppose that i's goal is that $C_j\varphi$ and that j cannot consistently deny the commitment. If i only extracts from j a move m that $C_j\varphi$, j has a winning strategy for denying i victory. She simply denies committing to φ (*I never said that*), since $C_j\neg C_j\varphi$ is consistent with $C_j\varphi$, even if $Bel_jC_j\varphi$. Player j lies, but she is consistent. If i manages to achieve $C_jC_j\varphi$, j can still similarly counter i maintaining consistency. *Only if* i achieves the *common commitment* $C^*_GC_j\varphi$, with G the group of conversational participants) does j not have a way of denying her commitment without becoming inconsistent, as $C^*C_j\varphi \rightarrow (C_jC_j\varphi \wedge C_jC_jC_j\varphi \wedge \cdots)$.

Speakers can not only deny prior commitments but also "undo" or "erase" them with *self-corrections*. For instance, if in (1)b 1 commits to not having a bank account; in (1)d 1 no longer has this commitment (See [G] for a detailed account of repair). Conversational goals of the form $C^*_GC_ip$ are unstable if i may correct herself; they may be satisfied on one

finite sequence but not by all its continuations. j's being able to correct a previous turn's commitments increases the complexity of i's goals, which affects the existence of a winning strategy for i; an unbounded number of correction moves will make any stable $C_G^* C_i p$ goal unattainable, if p is not a tautology. We observe, however, a sequence of self-corrections is only a good strategy for achieving j's conversational goals if she is prepared to provide an explanation for her shift in commitments (and such explanations must come to an end). As [V] argues, conversationalists are constrained to be credible in a certain sense if they are to achieve their conversational goals. Constantly shifting one's commitments with self-corrections leads to non-credibility, thus avoiding the problem of unbounded erasures.

We sketch a formalization of acknowledgments with two distinct dynamics of public commitments in a propositional language with two actions, one symbolizing the performance of a new utterance, and the other an acknowledgment of a previous turn. Let PROP denote a set of propositional variables and I a set of agents. Define the sets of actions and formulae as $\mathcal{A} := \varphi!^i \mid Ack^i(\alpha)$ and $\mathcal{L}_0 := \text{PROP} \mid C_i\varphi \mid [\alpha]\varphi \mid \varphi \to \psi \mid \bot$ where $\varphi, \psi \in \mathcal{L}_0$, $i \in I$ and $\alpha \in \mathcal{A}$. Formulas receive a Kripke semantics, and commitments by i are modeled via an accessibility relation for i. The semantics of formulae without action terms is as usual; e.g., $\langle \mathcal{M}, w\rangle \models C_i\varphi$ iff φ is true at every world accessible from w by the relation R_i. The truth of a formula $[\alpha]\,\varphi$ in a pointed model $\langle \mathcal{M}, w\rangle$ is obtained by checking the truth of φ against the pointed model *updated* by the action α, i.e. $\langle \mathcal{M}, w\rangle \models [\alpha]\varphi$ iff $\langle \mathcal{M}, w\rangle^\alpha \models \varphi$. We give two versions of this update operation yielding two different dynamics of commitment. The first one implements actions $\varphi!^i$ as restricting the commitments of x at w to $\varphi-$worlds, leaving commitments of other agents unchanged: for a world $v \in W$ let $R_i^{\mathcal{M}}(v)$ denote the set of worlds accessible from v by R_i in $\mathcal{M}$, and let $|\varphi|^{\mathcal{M}}$ denote the set of φ-worlds. Define for an utterance-action $\alpha = \varphi!^i$ $\mathcal{M}^{\alpha,w}$ to be the model with $\mathcal{W}^{\alpha,w} = \{w\} \cup (R_i^{\mathcal{M}}(w) \cap |\varphi|^{\mathcal{M}} \times \{i\}) \cup (W \times \{-i\})$ ($\mathcal{W}^{\alpha,w}$ thus brings together a copy of the φ worlds that i can access in one step from w in $\mathcal{M}$ and a distinct copy of W) and accessibility relations defined as (a)$\forall (v,i) \in W^{\alpha,w}\, R_i(w,(v,i))$ and (b)$\forall k \in I\, \forall l \in \{i,-i\}\, R_k((v,l),(v',-i))$ iff $R_k^{\mathcal{M}}(v,v')$. Finally define the updated pointed model as $\langle \mathcal{M}, w\rangle^\alpha = \langle \mathcal{M}^{\alpha,w}, w\rangle$. $\langle \mathcal{M}, w\rangle^{\varphi!^i}$ verifies that i commits to φ in w but changes neither i's second order commitments (it does not enforce $C_iC_i\varphi$) nor anyone else's commitments. A second, alternative definition for action update ensures that α affects commitments at all levels for all participants such that $[\varphi!^i]C_G^* C_i\varphi$. For an acknowledgment action $\beta = Ack^i(\alpha)$, the model is updated by applying the effects of action α in every world accessible for i in w. Formally, the set of worlds of the updated model is $W^{\beta,w} = W \bigcup_{v \in R_i^{\mathcal{M}}(w)} (W^{\alpha,v} \times \{v\})$ and the accessibility relations are: (a)$\forall v \in R_i^{\mathcal{M}}(w)\, R_i(w,(v,v)) \wedge \neg R_i(w,v)$ and (b)$\forall u,v \in W\, R_k(u,v)$ iff $R_k^{\mathcal{M}}(w,u) \wedge u = w \to k \neq i$ and (c) $R_k((u,u'),(v,v'))$ iff $u' = v'$ and $R^{\langle \mathcal{M},u'\rangle^\alpha}(u,v)$. On the second update definition, we can show that acknowledgments have no effect.

As for corrections, [LA] provide a *syntactic* notion of revision over the logical form of the discourse structure. Using the correction of m as an action update on the commitment slate prior to m yields a semantics for corrections. Our formal semantics captures the dynamic effects of announcements, corrections and acknowledgments; common commitments are important conversational goals and that particular conditions must obtain if they are to be achieved. In the full paper, we show how ambiguous messages affect the system.

Bibliography [LA] Lascarides & Asher, 2009, Agreement, Disputes and Commitment in Dialogue, *J. Semantics.* [AP] N. Asher & S. Paul, 2013: 'Infinite games with uncertain moves', *1st International Conference on Strategic Reasoning*, Rome. [Ca] T. Cachat et al., 2002: 'Solving Pushdown Games with a Σ_3 Winning condition' *Computer Science Logic.* [C] H. Clark *Using Language* 1996, CUP. [G] J. Ginzburg, *The Interactive Stance*, OUP, 2012. [T] D. Traum 1994, A computational theory of grounding in natural language conversations, Ph.D. thesis, Rochester. [V] A. Venant et al., 2014, 'Credibility and its attacks', *Semdial 2014.*